A Rákosi-éra pártjegyzőkönyveinek feldolgozása, elemzése és vizualizációja szövegalapú kapcsolatháló-elemzési módszerekkel

Gulyás Attila[1], Szabó Martina Katalin[1,2,3], Ifj. Boros István[1], Havadi Gergő[1]

[1]MTA TK „Lendület" RECENS Kutatócsoport
[2]Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3]Precognox Informatikai Kft.
{gulyas.attila, szabo.martina, boros.istvan, havadi.gergo}@tk.mta.hu

Kivonat: A jelen dolgozatban a hatalmi hálózatok szöveg alapú feltárását célzó projektünk egy részletét, a Rákosi-jegyzőkönyvek feldolgozását és az ezek alapján elkészített mintavizualizációt mutatunk be. A Rákosi-éra hatalmi hálózatának fejlődése és működése jól rekonstruálható a történelmi dokumentumoknak − levéltári anyagoknak, jegyzőkönyveknek − és interjúknak köszönhetően. Kutatásunkat a hatalmi hálózat mögött rejlő informális kapcsolatok feltárását célozza meg. Vizsgálati anyagaink között változatos forrásból származó, nagy mennyiségű irat szerepel, amelyek döntő többségükben különböző pártbizottságok ülését megörökítő jegyzőkönyvek. Ezen dokumentumok feldolgozása komoly kihívást jelent, mivel a számuk igen nagy, a digitalizált anyagok minősége pedig a közel használhatatlantól a jól olvashatóig terjed. Dolgozatunkban bemutatjuk a dokumentumok feldolgozásának az alapelveit, eszközeit és módszertanát, illetve ismertetjük, hogyan állíthatóak elő kapcsolathálók ezekből a dokumentumokból, majd tárgyaljuk a szövegalapú kapcsolatháló-elemzés módszerét, és betekintést nyújtunk az ehhez tartozó vizualizációs technikákba. Végezetül egy mintaelemzéssel szemléltetjük az alkalmazott módszert, megmutatjuk a szövegalapú kapcsolatháló-elemzésben rejlő lehetőségeket. Kutatásunk egyik fő célja az, hogy a szövegeken szentiment- és topikelemzést hajtsunk végre a jövőben, megerősítve, vagy éppen megcáfolva korábbi eredményeinket.

1 Bevezetés

1.1 A kutatás történelmi háttere

A második világháborút követően hazánkban kialakult politikai helyzetből 1949-re az MDP (Magyar Dolgozók Pártja) került ki győztesen.[1] Az ezt követő években egy szűk

[1] A Magyar Dolgozók Pártja 1948 júniusában jött létre, miután a Magyar Kommunista Párt „egyesült" a szociáldemokrata párttal (SzDP). Ezt hivatalosan fúziónak, vagyis a két munkáspárt egyesülésének nevezték, valójában azonban a meggyengített és megtizedelt szociáldemokrata párt maradványát olvasztotta magába az MKP (Magyar Kommunista Párt).

hatalmi elit a párthierarchia mellett a kapcsolatrendszerén keresztül biztosította uralmát. Kapcsolatrendszer alatt ugyanakkor nem csupán a politikai életben kiépített kapcsolatokról beszélünk, hanem az azon kívül, az informális életben zajló kapcsolatokról is. A későbbiekben is számos példát láthatunk arra, hogy pártfunkcionáriusok éppen informális kapcsolataikkal erősítették meg a politikai kapcsolataikat, vagy éppen az informális kapcsolataikból kovácsoltak politikai tőkét [1]–[3].

Kutatásunkban a politikai kooperáció során létrejött kapcsolatokat vetjük össze a párthierarchia által diktált struktúrával a kapcsolatháló-elemzés eszközeit segítségül hívva. A kutatás felfogható egyfajta történelmi elitkutatásnak, amelyet társadalomtudományos eszközökkel (a hálózatkutatás módszerével) végzünk, a történeti források prozopográfiai feldolgozásán és vizsgálatán keresztül (néhány ígéretes hazai kísérlet erre, például [4], [5].

Az elit Andorka alapján a társadalmi hierarchia csúcsán elhelyezkedő kis létszámú – az uralkodó osztálynál kisebb – csoport [6]. A politikai elit ennek egyik szegmense, típusa; a sztálinista típusú kommunista diktatúrákban az erőforrások (gazdaság, kultúra, társadalom és kapcsolódó tőkék) felett kizárólag regnáló csoport.[2]

A Rákosi éra hatalmi / politikai elitje szervezetileg viszonyalag könnyem körülhatárolható és definiálható: a Párt (MDP) politikai vezető testületeinek a Titkárság, a Politikai Bizottság, valamint az 1953-ig létező Szervező Bizottság) tagjaiból és póttagjaiból állt össze. Ezen belül is elsősorban (informálisan is) azok a csúcspolitikusok tartoztak bele megkérdőjelezhetetlenül és maradandóan a hatalmi elitbe, akik kiemelt pozíciókkal, személyes hatással és kapcsolatrendszerrel rendelkeztek (és ennél fogva információkkal bírtak) az államigazgatásban, a tömegszervezetek (pl. a szakszervezetek), kultúrális élet irányításában (a *Szabad Nép* főszerkesztője) avagy az erőszak szervezetek (ÁVH, Honvédség, Rendőrség) ellenőrzésében. Őket nevezhetjük a „a pártvezetés legfelső körének”.

Jellemző a Rákosi korszak hatalmi elitjére, hogy épp oly könnyen eshetett ki közülük valaki, ahogy bekerült (lásd. erről a folyamatos éberség hisztéria és ellenségkép fenntartását igazoló, Sztálini (szovjet) mintájú koncepciós perek sorát: Rajk, Marosán, Kádár, Kállai etc.).

Ebből a szempontból a későbbi Kádár alatti elit réteg lényegében a Rákosi alatti elitre épült, annak is gyorsan mozgosított, új másod- és harmad vonalára (Apró, Gáspár, Hegedüs, Komócsin, Münnich, Piros, Szalai, Vég vagy Czinege). Különösen érdekes lehet tehát, hogy ebből az elitből vezető pártelit mellett az említettek milyen, a párthierarchiában értelmezhető, de nem abból fakadó kapcsolatrendszert építhettek ki.

A MDP elnöke ugyan Szakasits Árpád lett, ám tényleges vezetője Rákosi Mátyás főtitkár volt. A párt létszáma meghaladta az egymillió főt. 1956 júliusában Rákosit leváltották az MDP éléről, utódjává a tőle politikájában nem sokban különböző Gerő Ernőt választották, akit október 25-én Kádár János váltott fel.

[2] Abban a közelmúlt történelmének kutatásával foglalkozó szakemberek szinte egybehangzóan egyetértenek, hogy nem érték- vagy presztízsalapon, hanem pozíciók vizsgálata mentén van leginkább értelme a pártállami elitbesorolásoknak illetve a hatalmi elit vizsgálatának (Rácz 2014).

1.2 Röviden a szövegalapú kapcsolatháló-elemzésről

A hálózatelemzés jelentőségére mutatnak rá azok az újabb, és egyre szaporodó kutatási eredmények, amelyek a legkülönfélébb hálózatos szerveződésekben szabályszerű mintázatok létrejöttére mutatnak rá. Barabási a következőképpen fogalmaz: „Hálózatok mindenhol vannak. Az agy axonok által összekötött idegsejtek hálózata, maguk a sejtek pedig biokémiai reakciók által összekötött molekulák hálózatai. A társadalmak szintén hálózatok […]. A hálózatok átjárják a technológiát is: az internet, az elektromos hálózatok, valamint a szállítási rendszerek csupán néhány példa erre.” [7]

A 1990-es évek végén jelennek meg az első olyan tudományos megállapítások, amelyek az eddig egymástól függetlennek tűnő hálózatos rendszereknek (pl. úthálózat, világháló, emberi kapcsolatrendszerek stb.) közös tulajdonságaira irányítják a figyelmet, továbbá amellett érvelnek, hogy ezek a tulajdonságok matematikailag leírhatók és elemezhetők [8]–[11].

A szövegalapú kapcsolatháló-elemzés alatt a szövegeket kapcsolathálóként értelmező és a társadalmi kapcsolathálók elemzési módszereit (SNA) használó paradigmát értjük [12]. Az SNA módszere a matematikából jól ismert gráfelméleti gyökerek mellett a fizika gyakorlatiasabb megközelítése alapján fejlődött ki [13].

A kapcsolatháló-elmélet és ezen belül a társadalmi kapcsolathálók elemzése hazánkban a köztudatba Barabási nyomán robbant be [14]. A szociológián belül a módszer leginkább a kapcsolati tőke vizsgálatában uralkodó megközelítés [15]–[17]. A kapcsolatháló-elemzés népszerűsége egyrészt abból fakad, hogy a módszer „társadalmi beágyazottsággal” rendelkezik [18].

Azonban nem csak társadalmi, (személy alapú) kapcsolatok, hanem tulajdonképpen „minden ábrázolható hálózatként” [12]. „Még a nyelv is, amit gondolataink közvetítésre használunk, önmagában véve nem más, mint szintaktikai kapcsolatokkal összekötött szavak hálózata” [7]. Ennek megfelelően a szöveget reprezentáló kapcsolatháló csomópontjai nem egyes személyek, vagy személyek csoportja, hanem az egyes szövegrészek, legtöbbször szavak. Gyakori ugyanakkor az úgy nevezett biparias kapcsolatháló is, melyekben általában a detektált kulcsszavak vagy témák mellé személyeket társítanak [19]. Így lehetséges például tudományos publikációk elemzése alapján arra következtetni, hogy mely kutatóhoz mely terültek társíthatók. Ennek segítségével következtethetünk arra, hogy ki, hogyan és milyen terület felé tolja azt a közösséget, mely az eszköz vagy adott téma fejlesztésével, művelésével foglalkozik. A szavak közötti reláció, uniplex [20] feltétele a bizonyos szöveg tartományon belüli együttes előfordulás.

A szövegalapú kapcsolatháló-elemzés tehát nemcsak egy újfajta reprezentációját jelentheti az szövegeknek, hanem a mögöttes tartalmak megismerésében is segítséget nyújthat.

2 A kutatás kérdései és forrásai

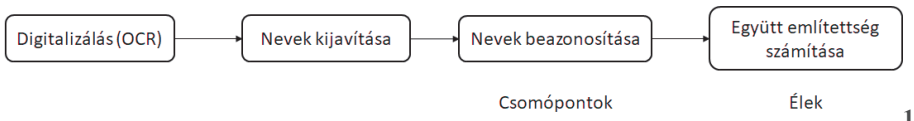
Kutatásunk célkitűzése a Rákosi-korszak (1948-1956) alatti politikai hatalmi elit látens hálózatának a felrajzolása, többfajta történeti forrás feldolgozása és elemzése segítségével. Vizsgáljuk a látens és manifeszt hierarchia, valamint a kapcsolatok dinamikájának történeti hálózatát.

Számos korábbi dolgozat feldolgozta már a pártelit látens kapcsolatainak alakulását: kiváló példákat olvashatunk az informális szférában kötött kapcsolatokról a közös vadászatokon keresztül [1], vagy említhetjük Aczél György erősen kapcsolathálókon alapuló, meglehetősen sikeres politikai működését [3]. Ennek nyomán kutatásunk egyik fő hipotézise az, hogy a politikai, vagy akár a politikai tevékenységeket kiegészítő elfoglaltságok terén együtt működő párttagok között olyan látens kapcsolat is létrejöhetett, amely a párthierarchiával párhuzamosan formálta a politikai szférában való tevékenységüket.

A forrásadataink az állampárt (MDP) politikai vezető testületeinek 1948-1956 között keletkezett üléseinek szerkesztett jegyzőkönyvei (Politikai Bizottság, Titkárság, illetve Szervező Bizottság). A hálózatban szereplő személyek pontos azonosításához egyéb történelmi dokumentumokat (biográfiákat, káderlapokat, és életrajzi adatbázisokat) használtunk fel, melyekből kiolvasható az adott személy politikai funkciója, és a pártéletben betöltött szerepe mellett számos további adat (iskolák, lakhelyek, különböző politikailag fontos eseményeken való részvétel stb.), amelyből ugyancsak informális kapcsolataikra következtethetünk. A rendelkezésünkre álló források tehát változó állapotban lévő gépelt, illetve gyakran kézi jegyzeteket is tartalmazó dokumentumok, amelyek feldolgozása és elemzése komoly kihívást jelent (részletesebben l. lentebb).

3 A feldolgozás és a kutatás módszertana

A korpusz létrehozását és feldolgozását az alábbi ábrának megfelelő legfontosabb lépésekben végeztük el:



1. ábra. A kapcsolathálózat létrehozásának a lépései

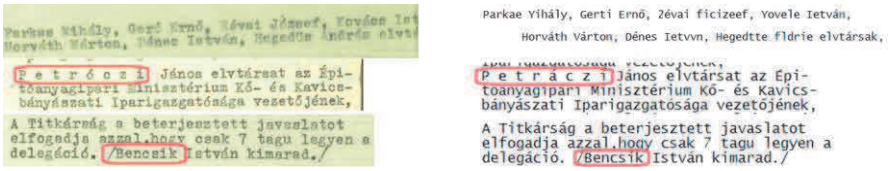
A jelen fejezetben e folyamatot részleteiben ismertetjük.

3.1 A korpuszszövegek feldolgozásának első lépései

Ahhoz, hogy a későbbi feldolgozási lépések anyagát létrehozhassuk, mindenekelőtt a szövegek digitalizálására volt szükség.

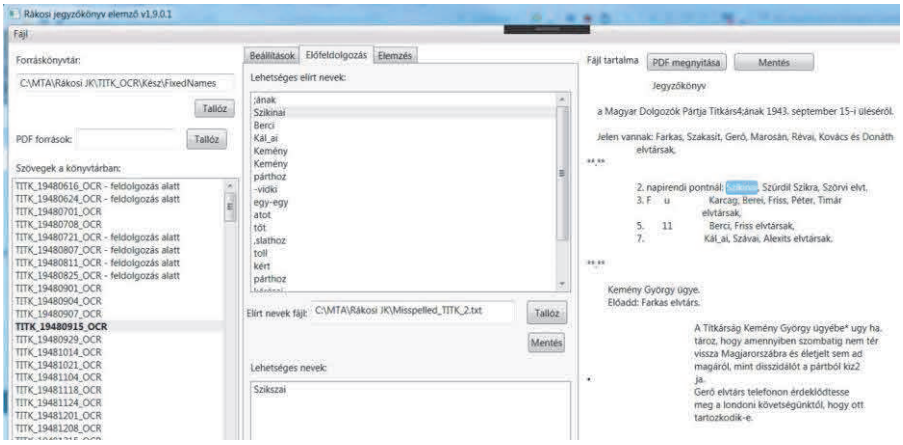
A digitalizálást OCR-eszközzel végeztük el. Ezt követően egy jelentős szövegkorrekciós fázist kellett beiktatnunk, elsősorban a további munkálatok szempontjából kardinális nyelvi elemek, a tulajdonnevek formai problémái miatt. A problémáknak több oka volt. Egyrészt, a szövegekben előforduló tulajdonnevek többsége az átalakítás során sérült, azaz különböző karakterhibák kerültek a szövegbe. Másrészt, a feldolgozott szövegekben találkozhatunk olyan írájelekkel is, amelyeket e történelmi dokumentumokban a nem történelmi szövegekhez képest sajátosan alkalmaztak, így például a per-jelet zárójel funkciójában, amelyet az OCR-eszköz automatikusan nem ismer fel, és ezért nem alakít is át zárójellé. Hasonló, a szöveg típusából fakadó egyedi jellemző az is, hogy bizonyos esetekben a tisztségviselők nevét úgy emelik ki, hogy a név minden egyes karaktere közé szóközt tesztnek („P e t r ó c z i"). Nyilvánvaló, hogy a nem történelmi szövegeken trénelt algoritmusok gyakorta nem az elvárásainknak megfelelően kezelik ezeket az egyedi megoldásokat. Végezetül, A nevekbe számos további okból kifolyólag is kerülhettek hibák a digitalizálás során. Így például adódhatott papír öregedéséből, az alkalmazott tinta gyenge minőségéből, vagy akár a digitálizálás során alkalmazott szkennelő eszköz nem kielégítő minőségéből is.

Az alábbi ábrán e típushibákra mutatunk néhány példát. A bal oldalon a forrás részletei, a jobb oldalon azok digitalizált verziói láthatóak.



2. ábra. (fentről lefelé) Rosszul digitalizált nevek; speciális név formátum; különleges konvenciók (a '/' karakter használata zárójelként)

Az OCR eszközzel végzett feldolgozást követően tehát a hibásan beolvasott nevek javítása volt az első lépés. Ehhez annak céljából, hogy az emberi erőforrás igényét csökkentsük, egyedi szoftverrel támogatott módszert használtunk, amely a rendelkezésre álló névtér-adatbázis alapján megkísérelte beazonosítani a szövegben fellelhető lehetséges neveket. Egy adott karaktersort akkor tekintett a program egy valós név lehetséges elírt alternatívájának, ha a karaktersor Levenshtein-távolsága [21] kevesebb volt a teljes név hosszának 30%-ánál. Így leginkább az OCR algoritmus által elkövetett típushibákat tudtuk javítani, például: Ger6 – Gerő, Kádár − K6d6r. A további hibákat manuálisan, az eredeti jegyzőkönyvekkel összevetve végeztük el. Ez utóbbiak azok az esetek voltak, amelyeket a szoftver azonosított ugyan, de akkora volt a torzulás mértéke (esetleg töredezett a név stb.), hogy azokat nem tekinthettük típushibának.



3. ábra. Kép a szoftverből

Az ezt követő lépés a nevek beazonosítása, amely tulajdonképpen a szövegekben előforduló névelemeknek az általuk jelölt entitásokhoz történő kapcsolását jelenti.

A tulajdonnevek azonosítása nem triviális feladat, ugyanis az egyes tulajdonnevek különböző alakokban is megjelenhetnek. A munka egy rendelkezésre álló névlista alapján történik, szoftverrel támogatott, ugyanakkor túlnyomórészt humán munkával. Ez a félautomatikus megoldás a következőképpen zajlik: Az azonosítást végző feldolgozók számára a szoftver minden egyes, a szövegben megtalálható név esetén felajánlja a lehetséges alternatívákat, kiegészítő információkat szolgáltatva az adott személyekről életrajzi adatok formájában.

A nevek azonosítását követően a szoftver az összes nevet a névtér-adabázis azonosítóját felhasználva taggé alakítja – például “Rákosi Mátyás”, vagy “Rákosi” helyére a “rakosi_matyas_8538” szót teszi a szövegben.

Azt a jelenséget, amikor ugyanarra az entitásra (személy, hely, szervezet stb.) különböző nyelvi jelölőkkel hivatkozhatunk, *koreferenciá*nak nevezzük [22]. A fentebb ismertetett munkánk eredményeképpen tulajdonképpen megtörténik a szöveg koreferenciaviszonyainak részbeni az azonosítása, azaz összekapcsoljuk azokat a nyelvi jelölőket, amelyek egyazon személyre referálnak [23], [24]. Azért hangsúlyozzuk, hogy az anonosítás csupán részbeni, mert az egyéb, nem tulajdonnévi alakú, de személyre referáló nyelvi elemeket (pl. névmások, E/3 ragozású igék stb.) nem vettük figyelembe, ugyanis arra nem volt szükségünk: ez a típus a jegyzőkönyvekben nem fordul elő, az egyes személyekre mindig tulajdonnévvel utalnak.

A tulajdonnevek azonosítása a további munkavégzés szempontájóbl rendkívül fontos lépés volt. Ez tette lehetővé ugyanis, hogy az együtt említettségeket megfelelően meg lehessen határozni, tehát azokat az eseteket is helyesen tudjuk kezelni, amikor teljes nevet, vagy pedig csak vezetéknevet említenek a jegyzőkönyvben.

A kapcsolatok elemzéséhez az első feladatunk a kapcsolathálók létrehozása volt, amelyekben az adott kapcsolatháló csomópontjait a személynevek, a köztük lévő

éleket pedig a közöttük detektálható valamilyen kapcsolat (kooperáció) megléte vagy hiánya adja.

Az együtt említést úgy definiáltuk, hogy a két egyedített jelölő egy adott bekezdésben 5 szó távolságon belül fordul elő. Fontos itt megjegyeznünk, hogy a digitalizálás miatt nem mindig lehetséges mondatok azonosítása, ezért muszáj a bekezdésre hagyatkoznunk, amelyek jól elkülönülnek. A munka jelenleg ennél fázisnál, azaz a nevek beazonosításánál és a kapcsolathálók létrehozásánál tart.

Mivel a jegyzőkönyvek egy jelentős részénél igen jó minőségű (80% feletti) az OCR minősége, lehetőségünk nyílhat egy más típusú és egyben részletesebb, szövegalapú kapcsolatháló-elemzés elvégzésére is.

3.2 Módszertan - a szövegstruktúra elemzése

A szövegalapú kapcsolatháló-elemzés egy nagyon fontos eredménye a szövegstruktúra olyan módon történő felrajzolása, amely rámutat a szövegben rejlő témákra (ez nem összetévesztendő a topikelemzéssel), és az ehhez tartozó szövegrészekre. Itt ezt a módszert mutatjuk be egy egyszerű példán keresztül.

A megoldás bemutatásához egy kis méretű tesztkorpuszt hoztunk létre hat darab véletlenszerűen kiválasztott Titkársági jegyzőkönyvből, melyeket igyekeztünk a lehető legteljesebb mértékben helyreállítani. Az így kapott anyag tehát a szövegfelismerésből származó hibákat nem tartalmazó, értelmes magyar nyelvű szöveg. Mivel a jegyzőkönyvek az elemzés szempontjából fontos egyedi struktúrával rendelkeztek, azt az eredeti dokumentumoknak megfelelő módon megőriztük. A kapcsolatháló-elemzést megelőzően a korpuszt a feldolgozáshoz szükséges tidy-text jellegű formátumra hoztuk, hogy a feldolgozó szoftverek számára elemezhető legyen.

Az itt bemutatott modell strukturális modell, mivel a szöveg teljes (a stopszavak nélküli) szókészlete (wordcount) megtalálható benne, és a fő célkitűzése a nagymennyiségű szöveg tartalmak feldolgozása a szöveg strukturájának vizualizálásával. A kapcsolathálóként vizualizált és kezelt szövegrészletek legyakoribb formája a szemantikus hálózatok (semantic networks).

A szemantikus hálózatok felépítése során rendszerint szótövezést (stemming, lemmatization) és N-grammokat alkalmaznak. Mivel a szövegalapú kapcsolatháló-elemzés a korpusz összes tartalmas szavának eredeti alakjára kíváncsi, nem alkalmazza a fenti eljárásokat, melynek pozitívumai és hátrányai egyaránt vannak. A nem információ hordozó szavak, másképpen a funkciószók (pl. a kötőszavak), valamint a jelen kutatásban vizsgált dokumentumokban, azok műfajából adódóan gyakran előforduló szavak (pl. *jegyzőkönyv*) szűrésére stoplistát alkalmaztunk, amelyet magunk állítottunk össze manuális módon a szövegek kézi elemzése alapján. A szavak közti kapcsolatot saját munkánk esetében is az együttes előfordulás adta. Bár az együttes előfordulás többfajta szövegegységen belül is értelmezhető, így lehet dokumentum, paragrafus, mondat, vagy egy bizonyos, előre meghatározott Δx szónyi távolság. A Δx szónyi távolság az adott szó szövegbeli pozíciójától mindkét irányban számított távolságot jelenti. Magunk ez utóbbi megoldást alkalmaztuk.

Jelen elemzésünkben csakis egy lexémából álló szavakat vizsgáltunk, szóösszetételeket és többszavas kifejezéseket nem vettünk figyelembe. Az általános

gyakorlattól eltérően nem végeztünk szótövezést, ugyanis az – amint azt a későbbiekben megmutatjuk (l. lentebb) – fontos információk elvesztését eredményezhette volna a számunkra.

Az így meghatározott, a kapcsolatháló alapjául szolgáló együtteselőfordulás-mátrixot a WORDij[3] [25] felhasználásával hoztuk létre. Végül, a kirajzolódó kapcsolatháló áttekinthetősége és a releváns kapcsolatok kiemelése érdekében csak azon szavak közé kerülnek élek, amelyek egy mondaton belül három szónál kisebb távolságon belül és legalább két alkalommal közösen előfordulnak.

A kapcsolathálózatok vizualizációinak az interpretálhatóságát jelentősen növeli, ha a pontok és az élek egyaránt színesek [26]. A kapcsolatháló megfelelő (bizonyos könnyen is objektívan interpretálható attribútumok szerinti) színezése nem csak az eredmények interpretációja mellett az elemzést is segíti. Az egyes közösségdetektáló algrotimusok megjelenítése, nagy méretű kapcsolathálók esetén nehezen kivitelezhető, és a szín atribútum hozzárendelése a közösségek detektálásnak fundamentális alapját képezi. A fentebb felsorolt fundamentális alapok kivetelezésének eszközéül a kapcsolatháló-elemzés és vizualizáció terén széles körűen elterjedt szoftvert a Gephi 0.9.1[4] [27] választottuk. E szoftver a vizualizációhoz kapcsolódó fejlett algoritmusokkal és grafikai, animációs képességekkel rendelkezik.

A kapcsolathálóban a pontok nagyságát a közöttiség központiság mutatója [12], [28], [29] adja, a színét pedig az, hogy mely kontextuális klaszterbe tartozik. Az élek akkor kapnak egyedi színezést, ha az általa összekötött pontok azonos klaszterbe tartoznak. Azon pontok kerülnek azonos klaszterekbe, melyek inkább összekötöttek, egy azonos nagyságú és sűrűségű véletlenszerű [13] gráf esetén várható élek számához képest. Az így előállt klasztereknek tehát tulajdonképpen a szövegben fellelhető "témák" feleltethetők meg.

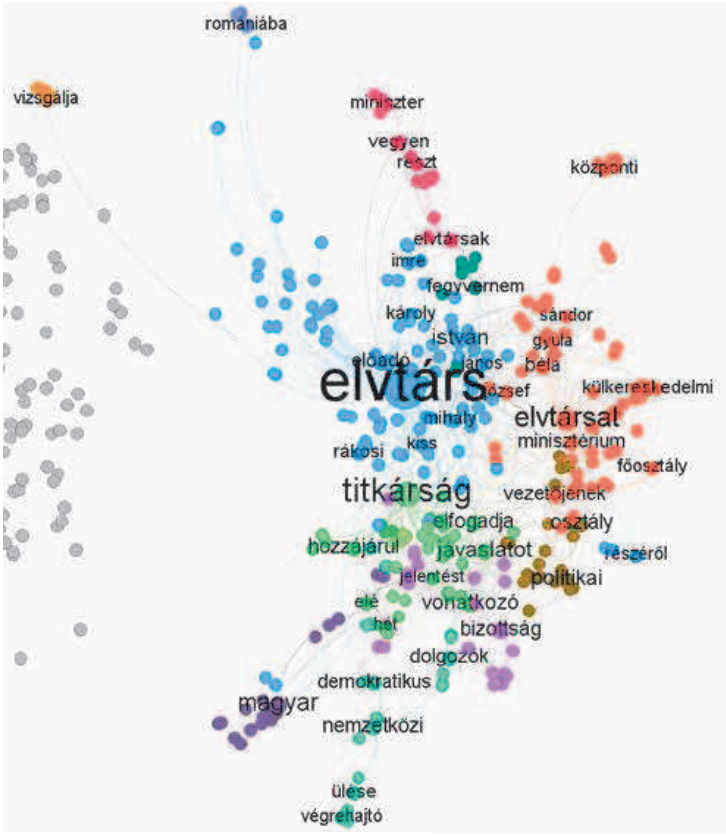
3.3 Eredmények

A továbbiakban a tesztkorpuszunkon végzett elemzés eredményét mutatjuk be. A célunk itt inkább a szövegalapú kapcsolatháló-elemzés bemutatása, mintsem az adott témával kapcsolatos hipotézisek tesztelése - ez már csak abból is fakad, hogy az előzőekben leírtaknak megfelelően egy a teljes forrásanyag állományhoz képest egy nagyon nagy mértékben leszűkített teszt korpuszt elemeztünk.

A korpuszt leképező kapcsolatháló egy 806 pontból (szavak száma) és 783 élből áll. Ebből az látszik, hogy a háló ritka, azonban a hálóban található élek inkább egy szűkebb csoportban találhatóak. Ezt a csoportot mutatjuk a 4. ábrán.

[3] http://wordij.net/

[4] https://gephi.org/



Személynevek, rangok, személyhez kapcsolódó utasítások, ügyek (13,4%)

Formaságok, a bürokratikus mechanizmus elemi (5%)

Utasítások, kinevezések, hozzárendelés (8,2%)

Az államapparátus és karhatalom elemei (2,9%)

Sajtóorgánumok, internacionális kapcsolatok, kulturális események (2,6%)

4. ábra. A tesztkorpuszt leképező kapcsolatháló

Ebben a pontok jelentős része 5-nél kevesebb kapcsolattal rendelkezik, s nagyon alacsony azon pontok száma, melyek negyvennél több kapcsolattal rendelkeznek. A pontok fokszámának átlaga 1.943, a legnagyobb előforduló fokszám 104. Ez alapján látható, hogy korpusz szókészletének varianciája alacsony és gyakori az azonos szópárok előfordulása.

A háló modularitása 0,65, ami azt mutatja, hogy a kapcsolathálóban azonosított klaszterek pontjai egy véletlenszerű gráf pontjainál jobban összekötöttek, tehát valós csoportokról beszélhetünk. A hálóban 420 közösséget találhatunk meg, amelyek közül öt darab nagyobb, mint a kapcsolatháló 2,5 %-a. Azok a pontok, melyek nem rendelkeznek kapcsolatokkal, magányos közösségeket alkotnak. Ezek nem láthatóak a fenti ábrán. Az azonosított témaklaszterek a teljes szöveg körülbelül egyharmadát tették ki.

Az elemzés során előálló klaszterek által lefedett témákat ugyancsak a 4. ábrán mutatjuk be. Látható, hogy a szöveg kapcsolathálóként való elemzése alapján jól elkülöníthetőek a fontos témák. Ez az eredmény már egy ilyen kis korpusznak az elemzésénél is fontos segítséget nyújthat, és egyúttal felhívja a figyelmünket egy a feldolgozással kapcsolatos érdekes jelenségre is.

Mint azt korábban említettük (l. fentebb), a feldolgozás során elhagytuk a szótövezést. Ennek jelentőségét az “Utasítások….” klaszter kiválóan mutatja, mivel ennek egyik központi szava az “elvtársat”, amely szót elveszítettünk volna, hogyha egy szótövezett korpuszon végezzük az elemzést.

4 Összefoglalás, további tervezett lépések

Dolgozatunkban ismertettük jelenleg futó kutatásunk nyelvtechnológiai szempontból is releváns munkálatait, valamint egy kisebb méretű korpuszon megmutattuk a szövegalapú kapcsolatháló-elemzés módszerének alkalmazását.

Kutatásunkban a Rákosi-korszakból származó pártgyűlési jegyzőkönyvek elemzésével foglalkozunk, célunk a párthierarchia mögött létező látens hierarchia felfedése a személyek közt zajló együttműködés, illetve bizonyos ügyekben való együttes érintettség vizsgálatával.

A feldolgozott és elemzett jegyzőkönyvekben a kapcsolatrendszert a nevek együttemlítettségével modelleztük.

A munkát több lépcsőben végeztük el. Mindenekelőtt, a digitalizált szöveg minősége szükségessé tette a szöveg korrekcióját, illetve a szövegekben lévő nevek korrekcióját. Kutatásunk jelenlegi fázisában az ezen korrekciót követő kapcsolatháló alkotást végezzük, amelyhez a jelentős élőmunka mellett komoly történelmi háttérismeretek szükségesek. Az ezt követő elemzéshez a szövegalapú kapcsolathálóelemzést használjuk fel, melyet egy kis méretű tesztkorpuszon mutattunk be. Megmutattuk, hogy a szövegeket kapcsolathálóként értelmezve úgy csoportosíthatóak a szövegben található szavak, hogy a csoportok elemzésével beazonosíthassuk a szöveg által lefedett témákat.

Az elmondottakon túl, ugyanennek az elemzésnek az eredményeként arra is rámutattunk, hogy az általános elemzési megközelítéssel ellentétben nem célravezető

az ilyen elemzést megelőzően szótövezést végezni, mivel az elfedheti adott szavak különleges funkcióit. Jelen eredményeink szerint ugyanis a szövegben azonosított témák egyike középpontjában egy toldalékkal ellátott szó áll.

A munka további lépéseként azt tervezzük, hogy az előfeldolgozott és névelem-azonosított szövegen szentiment- és emócióelemzést hajtunk végre. E feldolgozási lépésekhez a szótárillesztés módszert kívánjuk alkalmazni, amely például a gépi tanulás mellett egyszerűbb és költséghatékonyabb információkinyerési módszer [30]. A szentimentelemzéshez olyan lexikonra van szükségünk, amely a lexikai szinten pozitív vagy negatív értékelő tartalommal rendelkező nyelvi elemeket tartalmazza [31], [32]. Az emóciók felcímkézéséhez pedig egy olyan szótárra, amely a különböző érzelmek nyelvi realizációit tartalmazza [33], [34]. Nyilvánvaló, hogy a korpuszban feldolgozott szövegek sajátságai okán a kiinduló szótárak szöveganyagát majd jelentősen módosítani kell, valamint azok kiegészítésére lesz szükség.

E két tartalomelemzési megoldástól azt reméljük, hogy a segítségükkel további, a kapcsolathálózat szempontjából fontos szemantikai tartalmakat tárhatunk fel a jövőben.

5 Bibliográfia

[1] K. Bozsonyi, Z. Horváth, and Z. Kmetty, ‘A hatalom hálója - A Kádár-kori hatalmi elit hálózati struktúrája az együttvadászási szokások alapján’, *Korall*, no. 47, pp. 157–184, 2012.

[2] G. Majtényi, *K-vonal - Uralmi elit és luxus a szocializmusban - Uralmi elit és luxus a szocializmusban*. Nyitott Könyvműhely, 2010.

[3] E. Sík, ‘Aczélhálóban’, *Szociol. Szle.*, vol. 3, pp. 64–77, 2001.

[4] I. G. Kovács, *Elitek és iskolák, felekezetek és etnikumok - Társadalom- és kultúratörténeti tanulmányok*. Budapest: L’Harmattan, 2011.

[5] A. Rácz, ‘A budapesti hatalmi elit prozopográfiai vizsgálata 1956-1989’, Budapest, 19-Dec-2014.

[6] R. Andorka, *Bevezetés a szociológiába*. Budapest: Osiris, 2006.

[7] A.-L. Barabási, ‘A hálózatok tudománya: a társadalomtól a webig’, *Magy. Tud.*, no. 11, pp. 1298–1308, 2006.

[8] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Reprint edition. Cambridge: Cambridge University Press, 2008.

[9] L. Kovács, *Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban*. Tinta Könyvkiadó, 2013.

[10] D. J. Watts, *Small Worlds*. Princeton University Press, 1999.

[11] D. J. Watts, ‘The “New” Science of Networks’, *Annu. Rev. Sociol.*, vol. 30, no. 1, pp. 243–270, 2004.

[12] D. Paranyushkin, ‘Identifying the Pathways for Meaning Circulation using Text Network Analysis’, Oct-2011. [Online]. Available: http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/. [Accessed: 23-Nov-2017].

[13] P. Erdős and A. Rényi, ‘On Random Graphs I.’, *Publ. Math. Debr.*, vol. 6, pp. 290–297, 1959.

[14] Barabási Albert-László, *Behálózva - A hálózatok új tudománya*. Helikon, 2002.

[15] R. Angelusz and R. Tardos, ‘A gyenge kötések ereje és gyengesége’, in *Hálózatok, Stílusok, Kultúrák*, Budapest: ELTE Angelusz Róbert Társadalomtudományi Szakkollégium, 2012, pp. 101–127.

[16] M. Granovetter, ‘The Strength of Weak Ties’, *Am. J. Sociol.*, vol. 78, no. 6, pp. 1360–1380, May 1973.

[17] M. Granovetter, ‘A gyenge kötések ereje. A hálózatelmélet felülvizsgálata’, in *Társadalmak rejtett hálózata*, R. Tardos and R. Angelusz, Eds. Magyar Közvéleménykutató Intézet, 1991, pp. 371–400.

[18] R. Németh, ‘Módszerek a kvantitatív társadalomkutatási paradigmákban’, *SOCIO.HU*, vol. 3, no. 10.18030/SOCIO.HU.2014.3.27, pp. 1–42, 2014.

[19] M. Sedighi, ‘Using of co-word analysis method in mapping of the structure of scientific fields(case study: The field of Informetrics)’, *J. Inf. Process. Manag.*, vol. 30, no. 2, pp. 373–396, Feb. 2015.

[20] K. Takács, *Kapcsolatháló elemzés; Társadalmi kapcsolathálózatok elemzése|Digitális Tankönyvtár*. Budapest: Budapesti Corvinus Egyetem, 2011.

[21] V. I. Levenshtein, ‘Binary Codes Capable of Correcting Deletions, Insertions and Reversals’, *Sov. Phys. Dokl.*, vol. 10, p. 707, 1966.

[22] J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, ‘Coreference resolution: A review of general methodologies and applications in the clinical domain’, *J. Biomed. Inform.*, vol. 44, no. 6, pp. 1113–1122, 2011.

[23] E. Simon, ‘A magyar nyelvű tulajdonnév-felismerés módszerei’, Budapest, 2013.

[24] V. Vincze and R. Farkas, ‘Tulajdonnevek a számítógépes nyelvészetben’, in *Általános nyelvészeti tanulmányok XXIV.*, Akadémiai Kiadó, 2012, pp. 97–119.

[25] Danowski, J. A., ‘WORDij version 3.0: Semantic network analysis software’. University of Illinois at Chicago, 2013.

[26] L. C. Freeman and V. Duquenne, ‘A note on regular colorings of two mode data’, *Soc. Netw.*, vol. 15, no. 4, pp. 437–441, 1993.

[27] M. Bastian, S. Heymann, and M. Jacomy, ‘Gephi: an open source software for exploring and manipulating networks’, presented at the International AAAI Conference on Web and Social Media, 2009.

[28] U. Brandes, ‘A faster algorithm for betweenness centrality’, *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, Jun. 2001.

[29] L. C. Freeman, ‘A Set of Measures of Centrality Based on Betweenness’, *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[30] F. Drávucz and M. K. Szabó, ‘A beszélői szubjektivitás vizsgálata szentiment- és emóciókorpuszokon’, in *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből*, 2017, pp. 39–49.

[31] M. K. Szabó, ‘Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái’, in *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*, vol. 177, T. Gecső and C. Sárdi, Eds. 2015, pp. 278–285.

[32] M. K. Szabó, ‘A nyelvi értékelés mibenlétének kérdése a számítógépes értékeléselemzés (szentimentelemzés) szempontjából’, in *LingDok 15. Nyelvészdoktoranduszok dolgozatai*, Z. Gécseg, Ed. Szeged: Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola, 2016, pp. 153–172.

[33] M. K. Szabó and G. Morvay, ‘Emócióelemzés magyar nyelvű szövegeken’, in *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*, vol. 177, T. Gecső and C. Sárdi, Eds. 2015, pp. 278–285.

[34] M. K. Szabó, V. Vincze, and G. Morvay, ‘Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái’, in *Távlatok a mai magyar alkalmazott nyelvészetben*, Budapest: Tinta Könyvkiadó, 2016, p. 282−292.